Cocoa Price Prediction Based on Time Series Models*

Yizhuo Liu, Leo Kaixuan Cheng, Haobo Ren, Betty (Ruoran) Li

April-04-2024

Introduction

Cocoa is an important agricultural product, especially for some West African countries like Ghana and Côte d'Ivoire (). These two nations produce more than half of the world's cocoa supply. Therefore, changes in cocoa prices have very important effects on their economies and the income of farmers. This study aims to find the key factors that influence cocoa futures prices. The primary purpose is to develop forecasting models that help predict future price movements.

Cocoa prices often change due to external shocks. Climate change is one major factor. Studies show that abnormal weather can harm cocoa trees and reduce production. This supply shortage often causes prices to rise sharply. For example, in Ghana and Côte d'Ivoire, rising temperatures and irregular rainfall have already created serious problems for cocoa farmers ( ). Also, inflation and changes in exchange rates can change production costs and influence market demand. When inflation increases in cocoa-importing countries or the U.S. dollar strengthens, cocoa prices become more unstable ( ). Another important factor is the disease of the cocoa trees, which are susceptible to pests and viruses, such as the swollen shoot virus. These diseases reduce the number of healthy trees, leading prices to increase due to lower production ( ). All these factors are connected and must be considered together when modeling cocoa prices.

In this report, we will use time series forecasting models, including ARIMA and linear regression with exogenous variables. These models will help identify patterns of cocoa price predictions. The analysis is based on historical data and includes climate, economic, and agricultural indicators.

There are some key challenges in this study. The data may be missing or incomplete. Also, some time series are non-stationary, which means their trends change over time. Finally, combining different types of data can be very complex. To solve these problems, we will use careful preprocessing steps and test different models to choose the most accurate and reliable one.

Literature Review

Data

Methodology

We split our data into a training set and a test set using a 80% - 20% ratio. Besides the daily price data, we extract a monthly and yearly time series data by averaging all price in one month or one year to see if our models are robust enough to train on those data. For NA and missing values in the dataset, we interpolate the data directly.

*Code and data are available at: https://github.com/YizhuoLiu/ts_forecast_cocoa_price

Preliminary Model - ARIMA

Model Ensemble - ARIMA + Linear Model

ARIMAX

The dependence of the price on other factors such as weather and diseases may be important. Therefore, we choose to use a conditional ARIMA model which is called autoregressive integrated mean average with exogenous variables (ARIMAX). It takes care of the dependence of the time series price with other time series with an expression:

$$y_t = \Phi(B)(1-B)^d y_t = \Theta(B)\epsilon_t + \beta^T \curvearrowleft_t$$

where ϵ_t is a white noise sequence and x_t is our exogenous variable. If we expand this, we can get

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + \epsilon_t$$

This is a kind of conditional ARIMA, where our time series of interest is conditioned on other variables that are potentially influential. With that, we may be able to predict the price more accurately using the other variables. In practice, we use price as the y variable and we treat everything else in the dataset as the X variable. The ARIMA parameters are chosen based on inspection and model selection techniques such as AIC and MSE. The model is then fitted using a standard ARIMA technique.

Linear Regression on Lags

Temporal CNN

Convolutional neural network (CNN) is a popular machine learning model when dealing with image and sequential data. More precisely, when it comes to univariate time series data in our case, we will perform a 1d convolution, which basically convolves some kernels with the predefined receptive field before time step t. After that, take the loss function as that distance between the predicted time series and our training data then backpropagate the gradient as in usual machine learning practice. CNN can easily extract temporal dependence inherently. We are using dilation on CNN to better extract the feature.

More specifically, we are using an encoder-decoder structure. We encode the time series and the corresponding exogenous time series using temporal CNN into some hidden state. We treat different features of the time series as different channels. In this process, we also utilize recurrent neural networks (RNN) to make the network respect the long term dependency. The encoder consists of many residual blocks. Then, using multi-layer perceptrons (MLP), we construct a decoder to predict future values conditioned on the past time series as well as the exogenous variables. The model is fully differentiable so we optimize it using an AdamW optimizer with appropriate hyperparameters tuned by the validation error on a test set splitted from the dataset.

Results and Forecasting

Preliminary Model

Model Ensemble - ARIMA + Linear Model

ARIMAX

The ARIMAX model is fitted using daily data as well as the monthly data with parameter (10,2,1). The inference results on the monthly and daily data can be seen from the diagram below. We can see that the result is relatively bad although we take into consideration other factors. It is similar to the results with a vanilla ARIMA model: the predicted time series becomes straight as a line, the error bar is enormous compared to the scale of the time series. The overall upward trend is correct except for the drastically increasing. This shows that the model is very uncertain on its prediction, and simply cannot fit the data because of its complexity. This suggests that ARIMAX is not a good choice for this data.

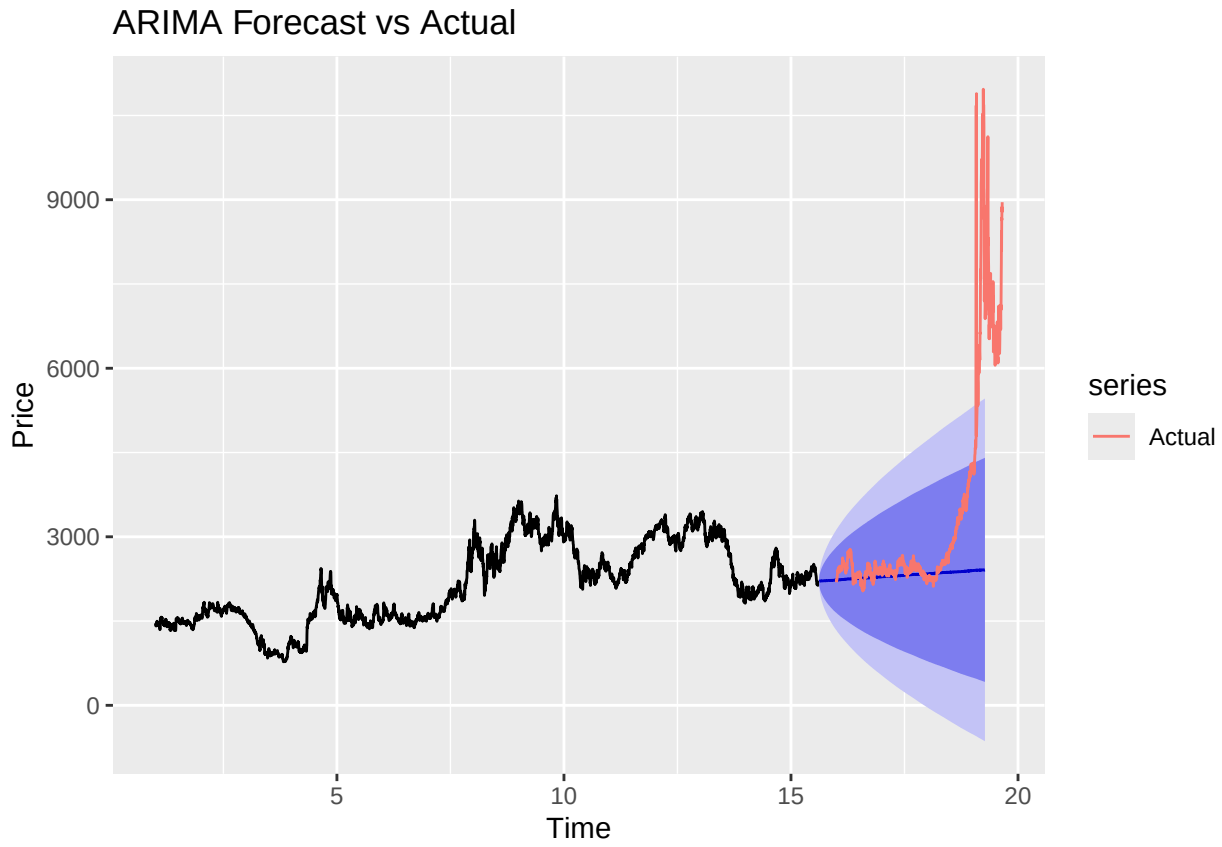


Figure 1: ARIMAX model fitted on daily data.

```
## Series: price_train
## Regression with ARIMA(6,2,1) errors
##
## Coefficients:
##          ar1      ar2      ar3     ar4      ar5     ar6      ma1  avg_PRCP
##       0.2528  -0.0548  -0.0217  0.0267  -0.0920  0.1555  -1.0000    2.4822
```

```
## s.e.  0.0597   0.0621   0.0607  0.0622   0.0621  0.0601   0.0108   19.6944
##       avg_TAVG  avg_TMAX  avg_TMIN
##         2.0623   -5.3472    4.1043
## s.e.    6.3268    4.2636    5.4348
##
## sigma^2 = 13372:  log likelihood = -1755.3
## AIC=3534.6   AICc=3535.75   BIC=3578.43
##
## Training set error measures:
##                       ME    RMSE      MAE          MPE     MAPE      MASE
## Training set -0.6758413 112.987 81.25517 -0.07094888 4.035911 0.2270426
##                       ACF1
## Training set 0.0009578105
```

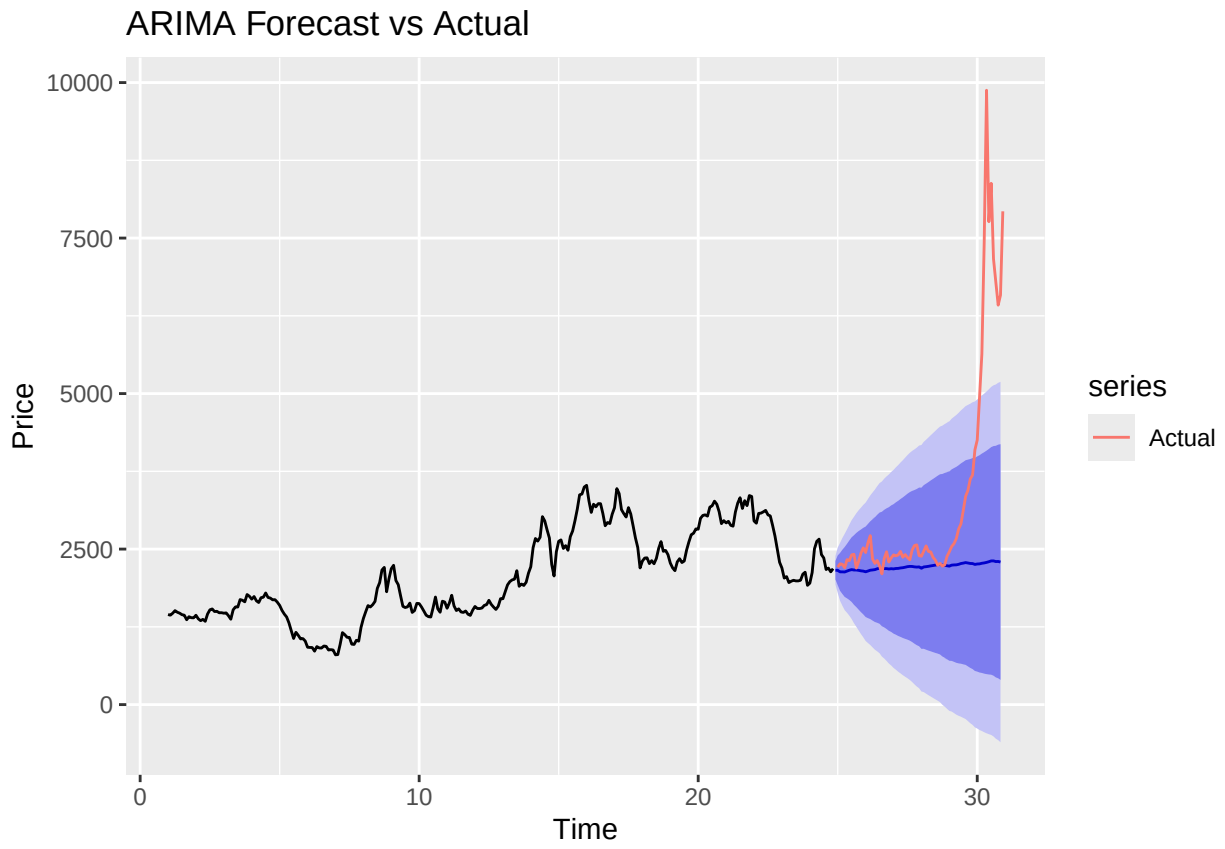


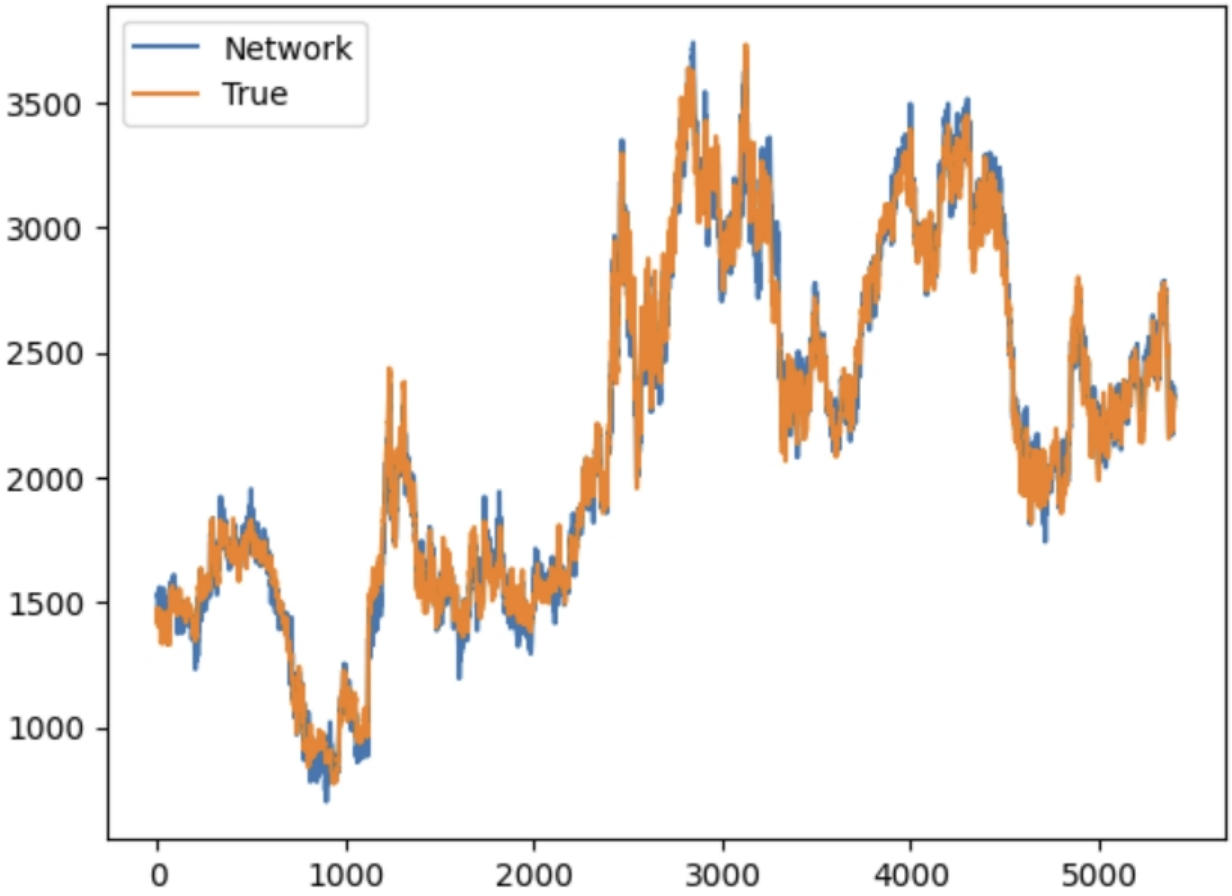
Figure 2: ARIMAX model fitted on monthly data.

Linear Regression on Lags

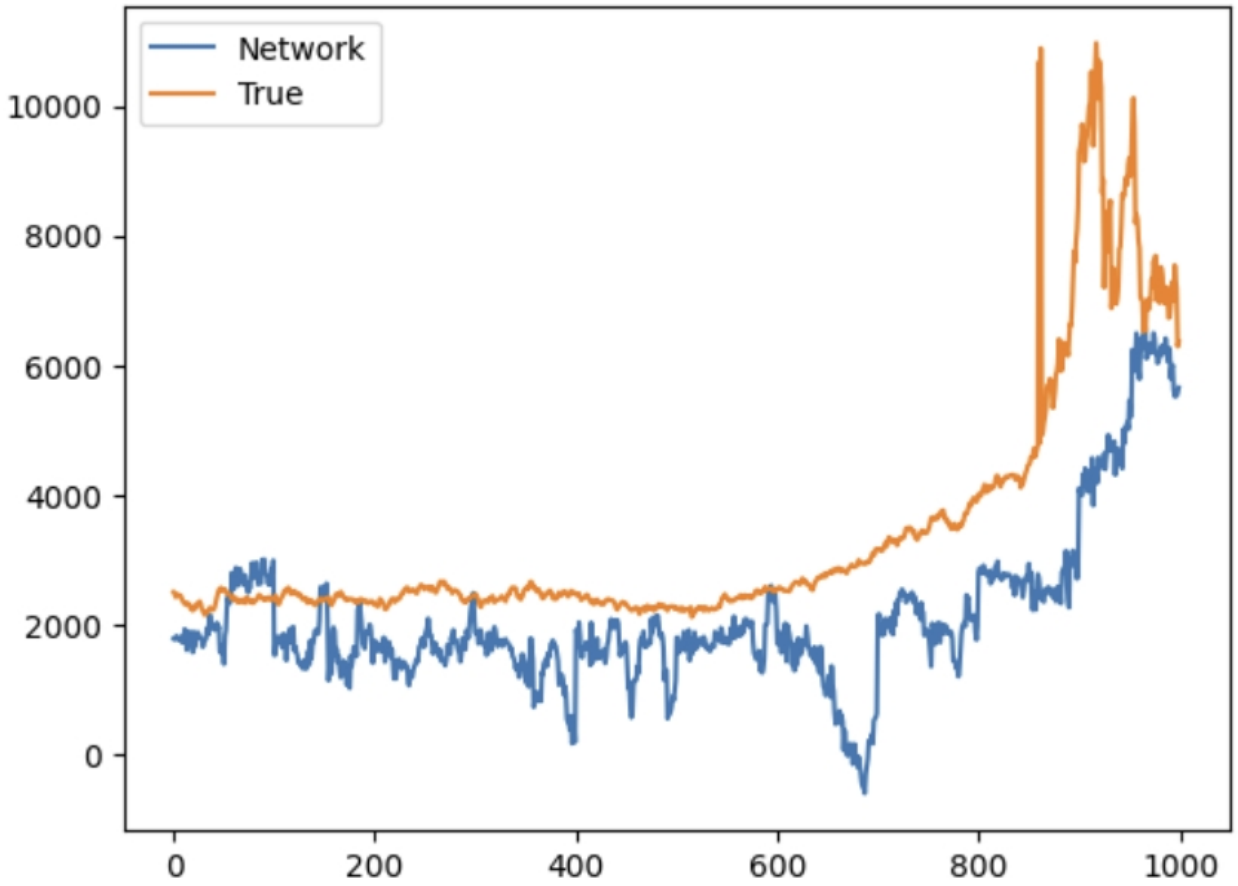
Temporal CNN

It takes the CNN ~1000 epochs to converge to a low loss. The fitted time series on the training and test set are shown. We can see that the model performs extremely well on the training set, showing that the expressiveness of the model is sufficient to capture the training data. We can see that the trend of the time series is correctly predicted by the CNN, but there is still some deviations from the true test set. The

validation MSE loss on the test set is also large. A potential reason is that the model is too complex and it overfits the given training set even though we apply a standard L^2 regularization on its weights. Overly complex model may not be suitable for a time series prediction without good regularization. The loss curve is also shown below.




```
{r myimage, echo=FALSE, fig.cap="The performance of the CNN model on train set. "} knitr::include_graphi
training set.png")
```



Discussion Overall the traditional time series prediction methods all perform relatively bad on the cocoa global price dataset. This shows that this data has an inherently complex nature, and many models are bad at capturing the pattern of the data. Another factor that may be problematic is the stationarity of the time series, which is hard to achieve. Also, traditional models such as ARIMA may fail because they don't take into consideration the exogenous variables such as weather and inflation. To resolve this, we found that the linear model trained on lags achieved a satisfying performance on both the training set and the generalization to the test set. Theoretically, temporal CNN will also perform well on this the performance of the linear model. However, due to issues like overfitting, the CNN model is not good as the linear model when predicting the future.

Although the vanilla ensemble model does not perform well, we may use the idea to combine two equally good models and tune the weight between the outputs of the two models to gain a better result. Or we can use another function to dynamically combine the results given by two or more models.

Data quality may also be important. We may incorporate more relevant data to train our models.

Recently there are works on utilizing diffusion models for time series prediction Yuan & Qiao (2024), which achieve state-of-the-art performance. Using that, we may be able to fit a robust and well-generalized model on this complex dataset conditioned on other factors. The flexibility of the proposed model enables us to do that.

Conclusion

Yuan, X., & Qiao, Y. (2024). *Diffusion-TS: Interpretable diffusion for general time series generation.* https://arxiv.org/abs/2403.01742